scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A near-telomere-to-telomere genome assembly of the Chinese soft-shelled turtle (*Pelodiscus sinensis*)

Zhe Li[1,3], Ruping Zhao[1,3], Zi'ao Wang[2], Lingrui Ge[2], Qin Qin[1], Shuting Xiong[1], Xiaoqing Wang[1] ✉ & Yazhou Hu[1] ✉

Pelodiscus sinensis **(Chinese soft-shelled turtle) is a freshwater species with economic and biomedical value. Here, we report a near-telomere-to-telomere (near-T2T) chromosome-level genome assembly. Using a combination of PacBio HiFi, Oxford Nanopore ultra-long reads, and Hi-C data, we assembled a 2.27 Gb genome (contig N50 = 132.52 Mb). Hi-C contact maps guided scaffolding and curation, anchoring 40 contigs to 34 chromosomes and resolving telomeres for 61 out of the 68 chromosomal ends. Gap filling and polishing yielded a chromosome-level assembly with no ambiguous bases (QV = 41.19; GC = 45.8%). BUSCO analysis indicated 97.9% completeness. Mapping rates for BGI short reads (99.4%), ONT long reads (100%), and HiFi reads (99.98%) confirm high accuracy across platforms. We annotated 21,124 protein-coding genes (97.77% functionally annotated) and 12,868 non-coding RNAs, with support from RNA-seq and full-length transcript data. Sex chromosomes (chrZ and chrW) were inferred from male-to-female depth ratios and validated using two known sex-linked SNP markers. This high-quality genome provides an essential resource for studying reptilian chromosome evolution, sex determination, and molecular breeding.**

Background & Summary

Pelodiscus sinensis, the Chinese soft-shelled turtle, is one of the most important freshwater aquaculture reptiles in East Asia[1]. It has attracted considerable scientific and commercial interest because of its high economic and medicinal value[2,3]. In China, *P. sinensis* is widely farmed and has become a major freshwater aquaculture species, with annual production reaching nearly 500,000 tonnes in 2023[4]. Its tissues are rich in essential amino acids (EAAs) and polyunsaturated fatty acids (PUFAs)[5,6]. In traditional Chinese medicine, various tissues of *P. sinensis* are used to treat ailments such as anemia, arthritis, and fatigue[7]. Recent biomedical studies have suggested that bioactive peptides derived from this species exhibit anti-inflammatory, anti-tumor, and cardioprotective activities. Therefore, *P. sinensis* is not only important for aquaculture but also a potential resource for the development of functional foods and pharmaceutical products[8–10].

Previous genome assemblies of *P. sinensis* have generally lacked high contiguity and completeness. The earliest assembly (PelSin_1.0), published in 2012, was limited to the scaffold level and contained numerous gaps and unresolved repetitive elements[11]. More recent assemblies, including rPelSin1.hap1 and rPelSin1.hap2, were generated under the Vertebrate Genomes Project (VGP) using PacBio HiFi and Hi-C technologies[12,13]. However, due to poor Hi-C data quality, these haplotype-resolved assemblies failed to meet VGP standards and were released as draft genomes without chromosome-level resolution or manual curation.

Here, we report the first near-telomere-to-telomere (near-T2T) genome assembly of *P. sinensis*, generated using a combination of PacBio HiFi long reads, Oxford Nanopore ultra-long reads, and Hi-C sequencing data. The final assembly spans 2.27 Gb with a contig N50 of 132.52 Mb, and includes 61 of the 68 expected telomeric regions. A total of 21,124 protein-coding genes were annotated, with 97.77% supported by transcriptomic and functional evidence. Moreover, sex chromosomes (Z and W) were identified based on sex-linked sequencing

[1]College of Fisheries, Hunan Agricultural University, Changsha, 410128, Hunan, China. [2]College of Animal Science and Technology, Hunan Biological and Electromechanical Polytechnic, Changsha, China. [3]These authors contributed equally: Zhe Li, Ruping Zhao. ✉e-mail: wangxiao8258@126.com; huyazhou@hunau.edu.cn

Fig. 1 Morphological characteristics of the Chinese soft-shelled turtle (*Pelodiscus sinensis*). (Left: ventral view; right: dorsal view).

Libraries	Clean reads number	Clean data (Gb)	Depth (×)	GC content (%)
HiFi	7,205,520	152.51	60.28	44.6
ONT ultra-long	706,445	70.42	27.83	—
Hi-C	722,958,697	215.57	85.21	43.96

Table 1. Summary of sequencing data. Note: Depth was calculated based on estimated genome size 2.53 Gb.

depth ratios and validated using two known sex-linked SNP markers. This high-quality genome provides a valuable resource for investigating reptilian genome architecture, microchromosome evolution, sex chromosome differentiation, and karyotypic innovations across amniotes.

Methods

Ethics considerations. All animal procedures were conducted in accordance with the institutional guidelines and approved protocol of Hunan Agricultural University and were reviewed and approved by the Animal Care and Use Committee of the College of Fisheries, Hunan Agricultural University (Ethics approval No. 2025 (209)).

Sample collection and sequencing procedures. A single adult female Chinese soft-shelled turtle (*Pelodiscus sinensis*) was collected from Changde City, Hunan Province, China (Fig. 1). Prior to sampling, the animal was anesthetized with tricaine methanesulfonate (MS-222) and then humanely euthanized under deep anesthesia according to the approved institutional protocol. Tissues were subsequently collected for sequencing. Liver tissue was sampled for genome sequencing. Additionally, tissues from the heart, spleen, kidney, intestine, brain, hypothalamus, pituitary, peripheral skin flap, gonad, liver, and muscle were collected for transcriptome sequencing to aid genome annotation. For genome survey sequencing, short-insert (300–500 bp) paired-end libraries were constructed and sequenced using the BGI DNBSEQ-T7 platform under a paired-end 150 bp (PE150) strategy. After quality control with SOAPnuke[14], a total of 141.94 Gb of clean data were obtained. Genome characteristics were estimated using GCE[15] (v1.0.2) with a 17-mer analysis, and the genome was estimated to be 2.53 Gb in size, with a heterozygosity rate of 0.42% and a repeat content of 59.23%. PacBio HiFi libraries (insert size 10–40 kb) were prepared using the SMRTbell Express Template Prep Kit v3.0 (Pacific Biosciences, USA) and sequenced on the PacBio Revio platform. This yielded 152.51 Gb of HiFi reads, with an average read length of 21,165 bp and an N50 of 22,828 bp (Table 1). To obtain ultra-long reads, genomic DNA was extracted using an SDS-based protocol and sequenced on the ONT PromethION platform using the SQK-ULK001 kit. A total of 70.42 Gb of ONT reads were obtained (N50 > 100 kb) (Table 1). Hi-C libraries were constructed using liver tissue. Chromatin was cross-linked, digested with MboI, biotin-labeled, and proximity-ligated before DNA purification. Libraries were sequenced on the BGI T7 platform, producing 221.54 Gb of raw Hi-C data. After filtering with Trimmomatic, 215.57 Gb of clean reads were retained (Q30 > 94% and a clean-read rate of 97.9%) (Table 1).

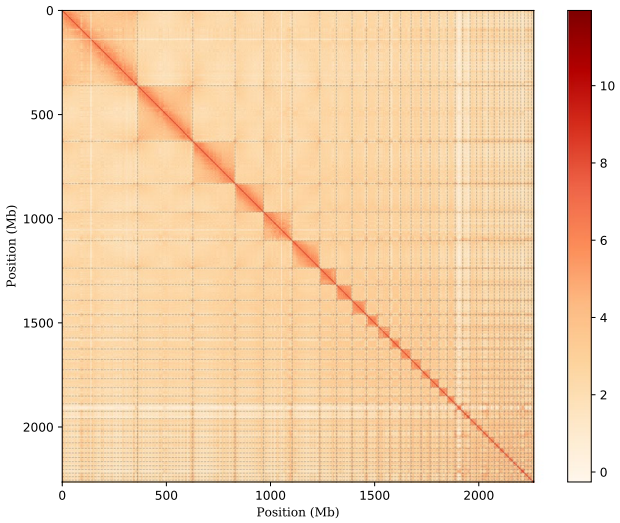

Fig. 2 Chromosomal Hi-C heatmap of the *P. sinensis* genome assembly. From left to right are chr1–chr16, chrZ, chr17–chr32, and chrW.

Term	Genes	Percentage (%)
Complete BUSCOs	3284	97.9
Complete and single-copy BUSCOs	3249	96.9
Complete and duplicated BUSCOs	35	1
Fragmented BUSCOs	29	0.9
Missing BUSCOs	41	1.2
Total BUSCO groups searched	3354	100

Table 2. BUSCO assessment results of the *P. sinensis* genome assembly.

Total RNA was isolated from a pool of the collected tissues. Sequencing on the BGI DNBSEQ-T7 platform yielded 13.12 Gb of clean RNA-seq data for gene structure annotation. To support this, we performed full-length transcriptome sequencing on the PacBio Sequel II platform, obtaining 50.36 Gb of subreads. From these, we derived approximately 1.03 Gb of high-quality circular consensus sequencing (CCS) reads, which were then processed to identify full-length non-chimeric (FLNC) transcripts for genome annotation.

Genome assembly and gap filling. To construct a high-quality reference, we adopted a hybrid assembly strategy to generate a near-telomere-to-telomere (near-T2T), gap-free chromosome-level genome of *P. sinensis*, integrating PacBio HiFi long reads, ONT ultra-long reads, and Hi-C sequencing data. The HiFi reads were assembled using Hifiasm (v0.19.9-r616)[16] in both the UL and the Hi-C integration modes, producing highly contiguous contigs. To enhance contiguity across complex repetitive regions, we performed independent de novo assembly of ONT reads using NextDenovo (v2.2)[17] with the parameters set as follows: read_cutoff = 1k and seed_cutoff = 25211, followed by six rounds of polishing with NextPolish (v1.3.0)[18]—including two rounds with ONT long reads and four rounds with short reads. The initial assembly spanned 2.30 Gb in 188 contigs, with a contig N50 of 132.52 Mb. Subsequently, Hi-C sequencing data were incorporated for chromosome-level scaffolding. Raw Hi-C reads were trimmed using Trimmomatic (v0.40)[19] and aligned with Juicer (v3)[20] to calculate contact maps, followed by misjoin correction and contig anchoring with 3D-DNA (v180922)[21]. After two rounds of scaffolding and manual curation using Juicebox (v1.11.08)[20], a total of 40 contigs were successfully anchored to 34 chromosomes, yielding a chromosome-level assembly (Fig. 2). The scaffolded assembly was 2.265 Gb in size, with a contig N50 and scaffold N50 of 132.52 Mb, and a chromosome anchoring rate of 98.0%. Gap filling was performed using TGS-GapCloser[22], which successfully closed the remaining inter-contig gaps based on ONT ultra-long reads. To further correct potential residual SNPs and structural variations (SVs), we applied the T2T-Polish pipeline[23]. The final assembly was 2,266,984,135 bp in length, with no ambiguous bases (Ns). We next evaluated the quality of this near-T2T, gap-free chromosome-level genome assembly in terms of its contiguity, completeness, and base-level accuracy. Mapping results showed high alignment rates of 99.4% (BGI), 100% (ONT), and 99.98% (HiFi), with high coverage and sequencing depth across platforms. BUSCO[24] analysis using the vertebrata_odb10 database found that 97.9% of the expected conserved genes were complete, indicating strong genome completeness (Table 2). The final genome had a QV score of 41.19 and a GC content of 45.8%, with no ambiguous bases detected.

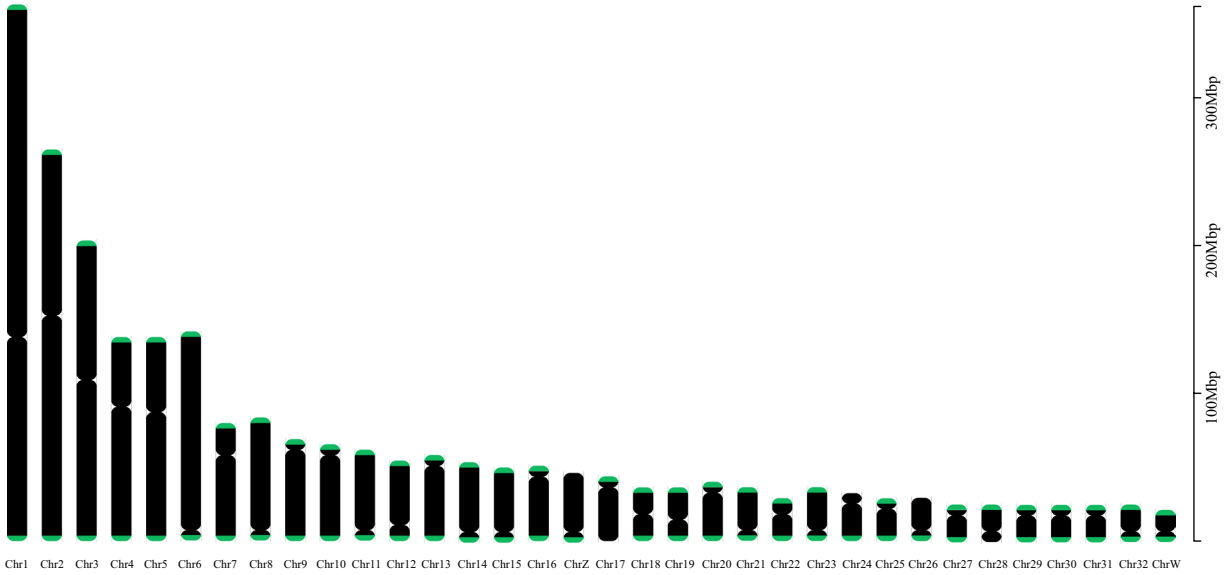

Fig. 3 Schematic diagram of centromeres and telomeres on chromosomes of the *P. sinensis* genome. Green indicates telomeres, and the indentations represent the positions of centromeres.

Type	RepeatMasker		RepeatProteinMask		*De novo* Length (bp)	*De novo* % in genome	Combined TEs Length (bp)	Combined TEs % in genome
	TEs Length (bp)	TEs % in genome	TEs Length (bp)	TEs % in genome				
DNA	206590142	9.11	7683887	0.34	211232250	9.32	339184196	14.96
LINE	283964060	12.53	187611948	8.28	701491004	30.94	757061324	33.40
LTR	142950556	6.31	32670964	1.44	214809927	9.48	281059042	12.40
Other	2631	0.00	0.00	0.00	0.00	0.00	2631	0.00
SINE	16957666	0.75	0.00	0.00	24548340	1.08	34085408	1.50
Unknown	8640867	0.38	0.00	0.00	33291938	1.47	40397457	1.78
Total TEs	663082438	29.25	227885268	10.05	1064856487	46.97	1285249687	56.69

Table 3. Transposable element statistics for the *P. sinensis* genome.

Telomere and centromere identification. For telomere identification, we screened the assembly for the canonical vertebrate telomeric repeat (CCCTAA)n using the VGP telomere-identification pipeline. For centromere identification, we used Tandem Repeats Finder[25] and pyTanFinder (https://github.com/Kirovez/pyTanFinder)[26]. From these, we selected candidate centromeric repeats that formed large, continuous tracts on each chromosome. The locations of the primary centromeric domains were then inferred by integrating these candidate repeat sequences with the interaction patterns observed in the Hi-C contact heatmap (Fig. 3).

Repeat annotation. Repetitive elements in the *P. sinensis* genome were annotated using a combination of homology-based and de novo approaches. First, known repeats were detected by aligning the genome against the RepBase database[27] using RepeatMasker (http://www.repeatmasker.org/) and RepeatProteinMask (http://www.repeatmasker.org/). Next, a species-specific de novo repeat library was constructed with RepeatModeler (http://www.repeatmasker.org/) and LTR-FINDER[28], and subsequently used for repeat annotation with RepeatMasker. In addition, tandem repeats were identified using Tandem Repeats Finder[25]. In total, 56.69% of the genome was annotated as repetitive sequences. Among these, LINEs represented the largest proportion (33.40%), followed by DNA transposons (14.96%) and LTRs (12.40%). SINEs and unknown elements accounted for 1.50% and 1.78%, respectively (Table 3).

Gene prediction and annotation. Gene structure prediction was performed using a combination of homology-based, transcriptome-based, and de novo strategies. For homology-based prediction, protein sequences from six closely related turtle species (*Pelochelys cantorii*[29], *Chelonia mydas*[30], *Chrysemys picta bellii*[31], *Emys orbicularis*[32], *Gopherus flavomarginatus*[33], and *Malaclemys terrapin pileata*[34]) were aligned to the *P. sinensis* genome using Exonerate[35]. Transcriptome-based prediction was conducted using RNA-seq data from eleven tissues and assembled with PASA[36] to support gene model construction. De novo gene prediction was carried out using AUGUSTUS[37] and GENSCAN[38]. All predictions were integrated into a non-redundant gene set using MAKER[39]. Functional annotation was performed by aligning predicted gene models to multiple databases, including NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), Swiss-Prot[40], Pfam[41], KEGG[42], GO[43], InterPro[44],

	Number	Percent (%)
Total	21,124	
InterPro	18,138	85.86
GO	14,268	67.54
KEGG_ALL	20,449	96.80
KEGG_KO	15,019	71.10
Swiss-Prot	15,994	75.71
TrEMBL	20,361	96.39
NR	20,623	97.63
Annotated	20,652	97.77
Unannotated	472	2.23

Table 4. Functional annotation statistics of protein-coding genes in *P. sinensis*.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		2,002	124.86	249,970	0.01
tRNA		9,488	76.15	722,501	0.03
rRNA	rRNA	145	1,506.50	218,442	0.01
18S	32	1,852.28	59,273	0.00	
28S	31	4,825.71	149,597	0.01	
5S	82	116.73	9,572	0.00	
snRNA	snRNA	1,233	164.97	203,413	0.01
CD-box	742	176.26	130,784	0.01	
HACA-box	105	139.81	14,680	0.00	
splicing	368	149.55	55,035	0.00	
scaRNA	15	181.87	2,728	0.00	

Table 5. Summary of non-coding RNA annotation in *P. sinensis*.

Term	Genes	Percentage (%)
Complete BUSCOs	3206	95.6
Complete and single-copy BUSCOs	3182	94.9
Complete and duplicated BUSCOs	24	0.7
Fragmented BUSCOs	70	2.1
Missing BUSCOs	78	2.3
Total BUSCO groups searched	3354	100

Table 6. BUSCO assessment results of the *P. sinensis* gene annotation.

and TrEMBL[45]. A total of 21,124 protein-coding genes were predicted. The average gene length was 52.03 kb and the average CDS length was 1.66 kb. Among these genes, 97.77% (20,652) were successfully annotated (Table 4). Non-coding RNAs (ncRNAs) in the *P. sinensis* genome were annotated using tRNAscan-SE[46], RNAmmer (v1.2)[47], and INFERNAL[48] based on Rfam models[49]. A total of 12,868 ncRNAs were identified, including 2,002 miRNAs, 9,488 tRNAs, 145 rRNAs, and 1,233 snRNAs, accounting for approximately 0.06% of the genome. The snRNAs included CD-box, HACA-box, splicing-related, and scaRNA subtypes (Table 5). BUSCO[24] analysis against vertebrata_odb10 revealed a completeness of 95.6% for the gene set, indicating a high level of completeness and annotation quality (Table 6).

Sex chromosome identification. To identify sex chromosomes in *P. sinensis*, we mapped short-read sequencing data from male and female individuals to the female near-T2T genome using BWA[50]. Quality control was performed with SOAPnuke[14]. The male-to-female (M:F) read depth ratio was calculated in 1 kb windows, and its distribution was plotted in R (v4.3.0). Chromosomes were classified based on this ratio: those with a ratio of approximately 1 were designated as autosomes, while those with a ratio of ~0.5 were assigned as the Z chromosome. The chromosome exhibiting a ratio approaching zero (indicating negligible coverage in males) was identified as the W chromosome (Figs. 4, 5). To further validate this result, we aligned two previously developed sex-linked SNP markers of *P. sinensis* to the genome using BLASTN[51]. The SNP loci were found to be precisely located on the candidate chrZ and chrW, consistent with their expected sex linkage, thereby supporting the reliability of the identified sex chromosomes (Table 7). Ultimately, a synteny circos diagram was constructed to depict the collinearity (Fig. 6).

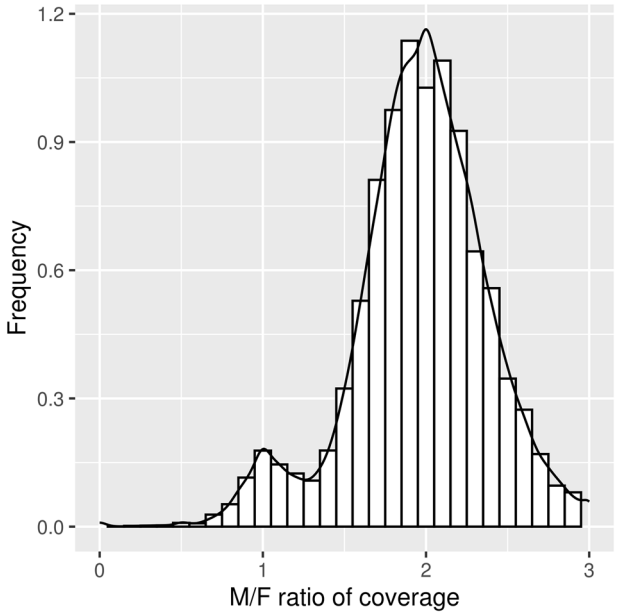

Fig. 4 Depth histogram of short-read alignments for male and female individuals on chrZ.

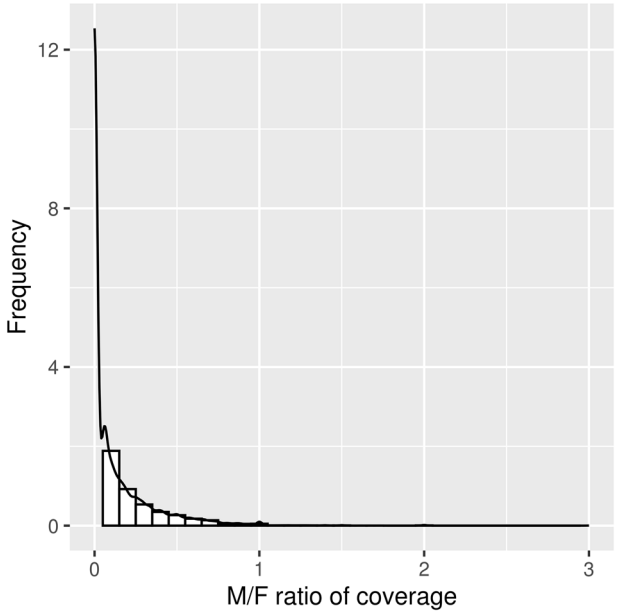

Fig. 5 Depth histogram of short-read alignments for male and female individuals on chrW.

Query	Subject	Identity (%)	Length (bp)	E-value	Position
SNP1	chrW	94.331	441	0.00	1171699-1171259
SNP1	chrZ	94.331	441	0.00	11616337-11616777
SNP2	chrW	97.291	443	0.00	11890177-11890619
SNP2	chrZ	96.652	448	0.00	39681705-39681258

Table 7. BLAST alignment results of sex-linked SNPs against chrW and chrZ.

Data Records

The near-T2T chromosome-level genome assembly of Pelodiscus sinensis is available under accessions GCA_052426305.1 (GenBank)[52], GWHGGEN00000000.1 (GWH)[53]. The gene annotation file (GFF) is hosted

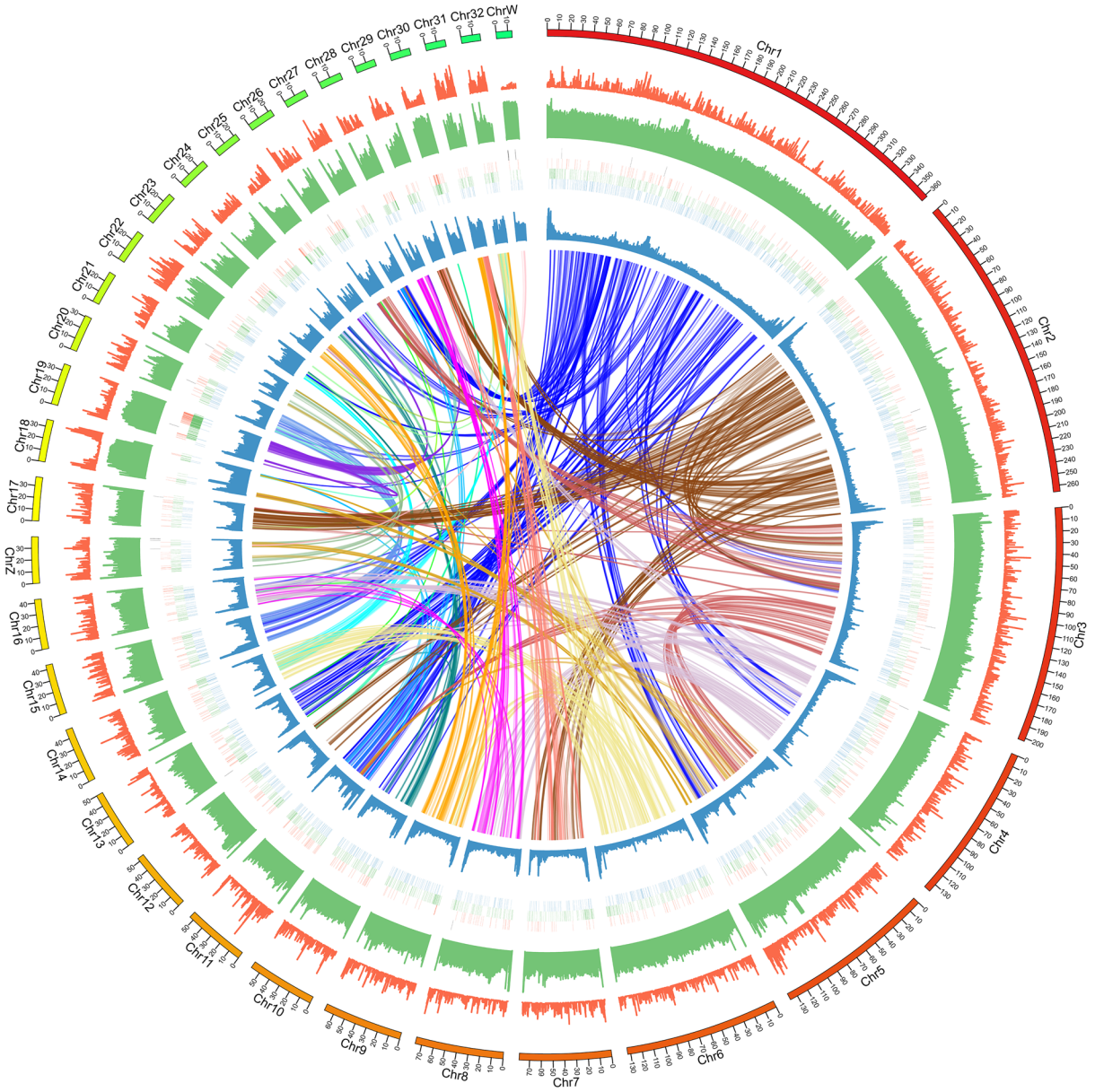

Fig. 6 Synteny circos diagram of the *P. sinensis* genome. From outer to inner circles: (1) Chromosomes, (2) Gene density, (3) Repeat count, (4) Non-coding RNAs (rRNA in black, snRNA in red, tRNA in green, miRNA in blue), (5) GC content, and (6) Collinearity.

at Figshare[54]. All supporting raw sequencing data are accessible at the Genome Sequence Archive (GSA) in the National Genomics Data Center under BioProject CRA032231[55].

Technical Validation

High-quality genomic DNA was extracted from liver tissue for sequencing library construction. The DNA samples exhibited a concentration of 114.5 ng/μl, an OD260/280 ratio of 1.82, an OD260/230 ratio of 2.12, and an average fragment size of 25 kb. The ratio of NanoDrop to Qubit concentration (N/Q) was 0.97, indicating reliable quantification and high purity. For transcriptome sequencing, RNA was extracted from pooled tissues including heart, spleen, kidney, intestine, brain, hypothalamus, pituitary, skin flap, gonad, liver, and muscle. The RNA samples met stringent quality requirements, with a total yield of 46.93 μg (sufficient for multiple library preparations), a concentration of 1042.9 ng/μl, a volume of 45 μl, an OD260/280 ratio of 1.99, an OD260/230 ratio of 1.00, and an RNA integrity number (RIN) of 6.8 as determined by an Agilent 2100 Bioanalyzer. Comprehensive quality assessments were conducted to evaluate the assembled genome of *P. sinensis*. The results demonstrated high completeness (97.9% BUSCO), base-level accuracy (QV 41.19), and structural integrity (high alignment rates and Hi-C contact maps), confirming the robustness of the final near-telomere-to-telomere genome assembly.

Data availability

The genome assembly is available in GenBank (accession GCA_052426305.1) and GWH (accession GWHGGEN00000000.1). The annotation file in GFF format is available at Figshare (https://figshare.com/s/503c57c9be8040e6a58d). The raw sequencing data (Hi-C, HiFi, and RNA-seq) have been deposited in the Genome Sequence Archive (GSA) under accession CRA032231.

Code availability

This study did not generate or use any custom code.

Received: 4 September 2025; Accepted: 18 December 2025;
Published online: 06 January 2026

References

1. Huo, L.-C., Liu, N.-Y., Wang, C.-J., Luo, Y. & Liu, J.-X. Lonicera japonica protects Pelodiscus sinensis by inhibiting the biofilm formation of Aeromonas hydrophila. *Appl Microbiol Biotechnol* **108**, 67 (2024).
2. Hou, M. *et al*. Comparative analysis for nutrients, flavor compounds, and lipidome revealed the edible value of pond-cultured male *Pelodiscus sinensis* with different ages. *Food Chemistry* **454**, 139795 (2024).
3. Bu, X., Liu, L. & Nie, L. Genetic diversity and population differentiation of the Chinese soft-shelled turtle (*Pelodiscus sinensis*) in three geographical populations. *Biochemical Systematics and Ecology* **54**, 279–284 (2014).
4. Ministry of Agriculture and Rural Affairs Bureau of Fisheries, P.R. China, National Fisheries Technology Extension Center of China & China Society of Fisheries. *China Fishery Statistical Yearbook*. (China Agriculture Press, Beijing,China, 2024).
5. Sun, C.-X. *et al*. Effects of fish meal replacement with animal protein blend on growth performance, nutrient digestibility and body composition of juvenile Chinese soft-shelled turtle *Pelodiscus sinensis*. *Aquaculture Nutrition* **22**, 315–325 (2016).
6. Liang, H. *et al*. Amino Acid and Fatty Acid Composition of Three Strains of Chinese Soft-Shelled Turtle (*Pelodiscus sinensis*). *Pakistan Journal of Zoology* **50**, 1061–1061 (2018).
7. Chu, J.-H., Chen, S.-M. & Huang, C.-H. Effect of dietary iron concentrations on growth, hematological parameters, and lipid peroxidation of soft-shelled turtles, *Pelodiscus sinensis*. *Aquaculture* **269**, 532–537 (2007).
8. Liao, P. *et al*. A novel ACE inhibitory peptide from *Pelodiscus sinensis* Wiegmann meat water-soluble protein hydrolysate. *Amino Acids* **56**, 40 (2024).
9. Pujiastuti, D. Y., Shih, Y.-H., Chen, W.-L., Sukoso & Hsu, J.-L. Screening of angiotensin-I converting enzyme inhibitory peptides derived from soft-shelled turtle yolk using two orthogonal bioassay-guided fractionations. *Journal of Functional Foods* **28**, 36–47 (2017).
10. Liu, L. *et al*. Studies on bioactive peptide from Chinese soft-shelled turtle (Pelodiscus sinensis) with functionalities of ACE inhibition and antioxidation. *African Journal of Biotechnology* **11**, 6723–6729 (2012).
11. Nguyen, S. N. Pelodiscus sinensis genome assembly PelSin_1.0. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_000230535.1 (2012).
12. Nguyen, S. N. Genome assembly Pelodiscus sinensis hap2. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_051311605.1 (2025).
13. Nguyen, S. N. Genome assembly Pelodiscus sinensis hap1. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_051311755.1 (2025).
14. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, gix120 (2018).
15. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology* **35**, 62–67 (2013).
16. Cheng, H. *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol* **40**, 1332–1335 (2022).
17. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol* **25**, 107 (2024).
18. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
19. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
20. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
21. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
22. Xu, M. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, giaa094 (2020).
23. Mc Cartney, A. M. *et al*. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat Methods* **19**, 687–695 (2022).
24. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
26. Kirov, I., Gilyok, M., Knyazev, A. & Fesenko, I. Pilot satellitome analysis of the model plant, Physcomitrellapatens, revealed a transcribed and high-copy IGS related tandem repeat. *Comp Cytogenet* **12**, 493–513 (2018).
27. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467 (2005).
28. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–268 (2007).
29. Liu, X. whole genome *Pelochelys cantorii*. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_032595735.1 (2023).
30. Mazzoni, C. Genome assembly *Chelonia mydas*. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_015237465.2 (2021).
31. Badenhorst, D. Genome assembly *Chrysemys picta bellii*. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_000241765.5 (2020).
32. Ciofi, C. Genome assembly *Emys orbicularis*. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_028017835.1 (2023).
33. Edwards, T. Genome assembly *Gopherus flavomarginatus*. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_025201925.1 (2022).
34. Wolak, M. Genome assembly *Malaclemys terrapin pileata*. *GenBank* https://www.ncbi.nlm.nih.gov/assembly/GCA_027887155.1 (2023).
35. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
36. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).

37. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–439 (2006).
38. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA1. *Journal of Molecular Biology* **268**, 78–94 (1997).
39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
40. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48 (2000).
41. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res* **42**, D222–230 (2014).
42. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
43. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
44. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354 (2021).
45. O'Donovan, C. *et al*. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* **3**, 275–284 (2002).
46. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).
47. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108 (2007).
48. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
49. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–124 (2005).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
52. Li, Z. Genome assembly *Pelodiscus sinensis*. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_052426305.1 (2025).
53. Li, Z. Genome assembly *Pelodiscus sinensis*. Genome Warehouse, China National Center for Bioinformation https://ngdc.cncb.ac.cn/gwh/Assembly/98803/show (2025).
54. Li, Z. Genome annotation of *Pelodiscus sinensis*. *Figshare* https://figshare.com/s/503c57c9be8040e6a58d (2025).
55. Li, Z. Raw data of *Pelodiscus sinensis*. *GSA* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRA032231 (2025).

Acknowledgements

This study was funded by National Natural Science Foundation of China (32102839), Natural Science Foundation of Hunan Provincial (2024JJ7550) and the earmarked fund for Hunan modern agricultural industry technology system (HARS-07), the Natural Science Foundation of Hunan Province (2024JJ8038).

Author contributions

Zhe Li and Ruping Zhao conducted the writing and data analysis. Zi'ao Wang, Lingrui Ge, Qin Qin and Shuting Xiong participated in data analysis. Xiaoqing Wang and Yazhou Hu provided guidance on experimental design and manuscript revision. Xiaoqing Wang and Yazhou Hu also provided project support.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.W. or Y.H.